DOCUMENT RESUME

ED 445 029		TM 031 634

AUTHOR	Deng, Hui; Ansley, Timothy N.
TITLE	Detecting Compensatory and Noncompensatory Multidimensionality Using DIMTEST.
PUB DATE	2000-04-00
NOTE	37p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Simulation; Test Construction; *Test Items
IDENTIFIERS	Compensatory Models; *DIMTEST (Computer Program); *Multidimensionality (Tests)

ABSTRACT

This study provided preliminary results about the performance of the DIMTEST statistical procedure for detecting multidimensionality with data simulated from both compensatory and noncompensatory models under a latent structure where all items in a test were influenced by the same two abilities. For the first case, data were simulated to reflect real test data in terms of descriptive statistics and classical item characteristics. In this case, DIMTEST did identify some degree of departure from essential unidimensionality for data sets from the noncompensatory model when the sample size was large and the interability correlation was low. For data simulated from the compensatory model, DIMTEST results suggested acceptance of the hypothesis of essential dimensionality. In the other three cases, data were simulated under various conditions in which the relative influence of the second dimension was greater than in Case 1. For these cases, when DIMTEST identified multidimensionality, the power increased when test length and sample size increased, and when interability correlation decreased. A question that remains unanswered is whether there are monotonic relationships between the power of DIMTEST and the degree of relative magnitude or variability for the two discrimination vectors. (Contains 8 tables and 11 references.) (SLD)

Detecting Compensatory and Noncompensatory Multidimensionality Using DIMTEST

by

Hui Deng and Timothy N. Ansley
The University of Iowa

Paper presented at the 2000 Annual Meeting of the
National Council on Measurement in Education, New Orleans, LA, USA.

Detecting Compensatory and Noncompensatory Multidimensionality Using DIMTEST

A fundamental assumption of most commonly used item response theory (IRT) models is that the test in question measures a single ability. This assumption must be evaluated before any application of unidimensional IRT models because violating this assumption could seriously bias item and ability parameter estimation (Ansley & Forsyth, 1985; Way, Ansley, & Forsyth, 1988). Therefore, it is crutial to verify this assumption prior to the use of unidimensional IRT models.

Among a variety of methods proposed to assess unidimensionality, DIMTEST (Stout, 1987, 1990) is a relatively new yet very promising statistical procedure that has attracted considerable attention in recent years. It was first developed by Stout (1987) and was further improved by Nandakumar and Stout (1993). This procedure is based on the conceptualization of *essential dimensionality*, which proposes to count only the dominant dimensions with minor dimensions ignored. This conceptualization depends on the replacement of local independence by the weaker notion of *essential independence*, and provides justification for the use of unidimensional IRT models subsequent to a statistical verification that essential unidimensionality holds for a set of item responses.

In order to apply DIMTEST, the N items of a test are split into three subtests: assessment subtests AT1and AT2, and partioning subtest PT. AT1 contains M items that can be selected either through factor analysis or expert opinion. These items presumably measure the same dominant ability and are dimensionally distinct from the rest of items. Another subset of M items are selected so that they have same item difficulty distribution as AT1 items. These items form AT2 and are used to offset the statistical bias in AT1 items arising from short test length and/or extreme difficulty levels. The remaining (N-2M) items form PT which is used to partition examinees into subgroups based on their total score on these items. The DIMTEST statistic T is computed for AT1 and AT2 subtests based on the within subgroup differences between the usual variance estimate and the unidimensional variance estimate. This statistic has been proven to be asymptotically normally distributed with mean zero and variance one when essential unidiemnsionality holds (Stout, 1987).

DIMTEST has many advantages such as its nonparametric nature, asymptotic theory basis, and computational efficiency. Its performance has been evaluated by many studies based on simulated and real data. Generally the results indicate that DIMTEST is able to correctly confirm unidimensionality for unidimensional datasets and effectively detect multidimensionality for two- or three-dimensional datasets (Nandakumar, 1993; Nandakumar & Stout, 1993; Stout, 1987). The accuracy of DIMTEST has been found to depend on both sample size and test length, with T performing best on tests with more than 25 items and with sample sizes greater than 500 (de Champlain & Gessaroli, 1991). Nandakumar & Stout (1993) also found that DIMTEST performed poorly when a test contained highly discriminating items with guessing present. Therefore they revised this procedure to overcome this limitation and automate the determination of M, the size of the assessment subtests. The improved DIMTEST, with statistic T', has been shown in simulation studies to adhere more closely to the nominal level of significance for unidimensional tests and achieve greater power for multidimensional tests (Nandakumar & Stout, 1993).

When simulated data are used to assess the sensitivity of DIMTEST to multidimensional data, one needs to choose a multidimenisonal IRT model as a basis for data generation. Does the choice of model impact the performance of DIMTEST? In other words, does DIMTEST distinguish multiple dimensions equally well with data generated from different multidimensional models? This is the major question that has driven the present study.

Among the multidimensional models that have been proposed, a major difference rests on whether compensation occurs among the abilities required to answer the items correctly. Both compensatory and noncompensatory models have been proposed. Sympson (1978) proposed a multidimensional extension of the unidimensional three-parameter logistic model that can be classified as noncompensatory (or partially compensatory). This model can be represented as

$$P_{ij}(\theta_{ih}) = c_j + \frac{1-c_j}{\prod_{h=1}^{n}\{1+\exp[-1.7a_{jh}(\theta_{ih}-b_{jh})]\}}, \quad (1)$$

wher θ_{ih} is the ability parameter for person i for dimension h, a_{jh} is the discrimination parameter for item j for dimension h, b_{jh} is the difficulty parameter for item j for dimension h, and c_j is the guessing parameter for item j.

A compensatory multidimensional extension of the three-parameter logistic model was represented by Doody-Bogan and Yen (1983) as

$$P_{ij}(\theta_{ih}) = c_j + \frac{1-c_j}{1+\exp[-1.7\sum_{h=1}^{n} a_{jh}(\theta_{ih}-b_{jh})]}, \quad (2)$$

where all parameters are defined as in equation (1).

The distinction between the two models can be intuitively seen by comparing the denominators of Equation (1) and (2). In the noncompensatory model, the denominator is the product of denominators for each dimension, while in the compensatory model the effects of each dimension are combined within the exponential term. Therefore the compensatory model permits high ability on one dimension to compensate for low ability on another dimension in terms of probability of correct response; whereas in the noncompensatory model high ability on one dimension cannot offset low ability on another dimension outside of a limited range, since the maximum probability of correct response based on one dimension is the upper bound for the probability based on the two dimensions.

In addition to model selection, another issue concerning simulating multidimensional data is the specification of a latent structure underlying test items. Based on a simple structure pattern, items of a test can be partitioned into clusters that are each influenced by a single ability. With a less-clear-cut latent structure, a test may contain some "mixed" items that are influenced by more than one dimension. In a more extreme situation, each item in a test can be influenced by the same multiple dimensions.

As outlined above, specifing model and latent structure are the two major issues to be considered when simulating multidimensional data. Data can be simulated in different ways depending on how choices are made on these two issues. A review of literature suggests that simulation studies for DIMTEST have almost uniformly used a compensatory model coupled with a simple structure to generate data. That the compensatory model has always been the choice is largely due to the fact that there is no estimation procedure currently available for the noncompensatory model; therefore, no parameter estimates from real data can be used for data simulation. This presents the question of how to make the simulated data realistic if a noncompensatory model is to be used for data generation. It has been argued that the noncompensatory view of dimensionality is more reasonable when a multidimensional test is considered to be one that requires the simultaneous application of two or more abilities (Ansley & Forsyth, 1985; Sympson, 1978). If this is the case, how well DIMTEST can distinguish multiple dimensions when data arise from a noncompensatory model needs to be studied, and the results compared to those from compensatory cases. For this kind of study, however, it is very important to ensure that the simulated data represent real test data as well as possible.

With respect to the specification of latent structure, a general approach shared by the previous studies is that, a test was taken to consist of a subset of items dependent on θ_1 alone, another subset of items dependent on θ_2 alone, and sometimes, a third subset of items dependent on both θ_1 and θ_2. However, such a simple structure type of pattern may not represent what one might typically encounter in real testing situations. With real data it is sometimes the case that all of items in a test are simultaneously influenced by the same multiple abilities. For example, a general reading ability may influence all of items in a math-problem solving test. As another example, a test anxiety factor may also influence all the items in a test instead of just a few. It seems necessary to simulate data to reflect this type of situation when investigating how well DIMTEST can detect multidimensionality.

In recognition of the importance of this type of latent structure, Nandakumar (1991) conducted a simulation study which included two cases of multidimensional structures. In one case several minor abilities existed with each influencing only a small group of items, while in the other case one minor ability existed which influenced all items in the test. In both cases a compensatory model was used for data simulation. The mean and standard deviation of item discrimination parameters for the minor ability were considered to reflect the influence of the minor ability relative to the major ability. A rough index β was further proposed to assess the deviation from essential unidimensionality due to the joint variation of a_1 and a_2, with the index defined as the minimum of the a_1 and a_2 variances multiplied by a constant. According to the results from that study, DIMTEST tended to retain the hypothesis of essential unidimensionality when the minor dimension(s) had a relatively small influence on item scores, and was more likely to reject the hypothesis when the influence of minor dimension(s) increased. The rejection rates were also shown to vary roughly according to β, with higher rejection rates associated with higher values of β. Among simulation studies for DIMTEST, this one is of special interest since it for the first time simulated data based on a new type of latent structure. However, this study only simulated cases with uncorrelated abilities and uncorrelated item parameters, which limits the extent to which the results can be generated to other conditions.

Another study by Hattie et. al is also noteworthy since it was the initial attempt to examine the performance of DIMTEST using data simulated from a noncompensatory model (Hattie, Krakowski, Rogers, & Swaminathan, 1996). In that study, data were generated using a program called DIMENSION based on a simple structure pattern. The effectiveness of DIMTEST for identifying compensatory and noncompensatory multidimensionality was examined along with some other issues. As a result, DIMTEST was found to be sensitive to whether the multidimensional data arose from a compensatory or a noncompensatory model. Specifically, for data from a compensatory model, the null hypothesis of essential unidimensionality was appropriately rejected most of the time, whereas for data from a noncompensatory model, the null hypothesis was rejected far less than expected under most conditions. Their conclusion was that DIMTEST is only applicable for identifying compensatory multidimensional data. However, this study has a major limitation in that it did not address the issue of realism for any of the simulated data sets, and the way the data were simulated may have been problematic, especially in the noncompensatory case. The concern is centered at the two difficulty ranges, [-2 , -1, 0, 1, 2] and [-1, -.5, 0, .5, 1], that were used for data simulation. As has been pointed out (Ansley & Forsyth, 1985), in the data generation procedure using a noncompensatory model, the difficulty (*b*) values play a major role in determining the realism of the data sets. It has been shown that data sets simulated from a noncompensatory model with *b* vectors centered at zero resulted in test data indicative of an uncharacteristically difficult test. Therefore the *b* values need to be scaled to have lower means to avoid such a problem.

None of the studies in the literature has examined the performance of DIMTEST for data simulated based on a noncompensatory model with a latent struture in which all items load on the same dimensions. This represents a situation where all the items of a test are influenced by the same abilities and all abilities are required simuataneously to answer each item correctly. Referring back to the example of math-problem solving test, the reading ability is considered to influence all the items in the test; in addition, the two abilities may not be compensatory. For an examinee very low on the major ability (math-problem solving), no degree of competence on the minor ability (reading) may be able to compensate for this deficiency and lead to high probability of correct response. This situation may be of much practical relevance and should be considered in simulation studies.

The purpose of this study, therefore, was to examine the power of DIMTEST for detecting multidimensionality with data simulated using both compensatory and noncompensatory models based on a latent structure in which each item is simultaneously influenced by the same two abilities. Different simulation cases were considered which varied in terms of relative potency of the second dimension, in order to gain knowledge of certain performance characteristics of DIMTEST under these conditions. The performance of DIMTEST for identifying multidimensionality was assessed and compared across the two models. The effects of test length, sample size, interability correlation, and guessing on the power of DIMTEST were also examined.

Method

Monte Carlo simulations were conducted for each simulation case which varied in terms of the distribution of item discrimination parameters. For the noncompensatory model, the item difficulty parameters determine, to a large extent, the realism of the generated datasets, therefore the item difficulty parameters were not altered across simulation cases. The relative dominance of the second dimension was manipulated by means of changing the distribution characteristics (mean, SD) of the item discrimination parameters.

Simulation Case 1

The distribution characteristics for the a and b vectors were adopted from the Way et. al (1988) study. Specifically, for the noncompensatory model, the a_1 values had a mean of 1.23 and a SD of 0.34, while the a_2 values were centered at 0.49 with a SD of 0.11. The two a vectors had a correlation of -0.29. The b_1 values had a mean of -0.33 and a SD of 0.82, while the mean and SD for b_2 values were -1.03 and .82. The correlation between the b vectors was .38. The c value was set at .2 for all items. The item parameters for the compensatory model were obtained by adding the following constants to the corresponding item parameters for the noncompensatory model: -.20 to each a_1 value, .63 to each b_1 value, and 1.0 to each b_2 value. The a_2 values and c values were unchanged. The rationale for the selected parameter distributions can be found in the two previous studies (Ansley & Forsyth, 1985; Way, Ansley, & Forsyth, 1988). These sets of item parameters have also been shown in these studies to yield item responses that closely resembled actual test data in terms of descriptive statistics, reliability, and difficulty indices; and the number-correct score distributions resulting from the two models were reasonably similar.

To generate binary item responses, the a_1 and a_2 values and the b_1 and b_2 values were each generated from a bivariate normal distribution with the corresponding means and SDs specified above. Examinee abilities were generated from a bivariate normal distribution with both means zero and both variances one and with a certain level of interability correlation (.3 or .7). For each simulated examinee, the probability of correctly answering each item was computed using either of the two models with the corresponding item parameters and the generated ability for the examinee. If a uniform random deviate in the interval (0, 1) was less than or equal to the computed probability, the examinee was considered to have answered the item correctly and was given a score of 1; otherwise a score of 0 was given.

The design of the study used three sample sizes (500, 1000, 2000). Each dataset was partitioned into two groups. One group (of size 200, 300, 500) was used for factor analysis to select the DIMTEST AT1 items, and the other group (of size 300, 700, 1500) was used to compute the statistic T.

In addition, three test lengths (20, 40, 50), two levels of interability correlation (.3, .7), and two choices of model (compensatory, noncompensatory) were used. All factors were completely crossed, resulting in a total of 36 combinations. Each combination was replicated 100 times for a total of 3600 datasets, with new examinee responses being simulated each time. DIMTEST was applied to each dataset. For all DIMTEST runs, the default method of factor analysis was used for selecting AT1 items, and the Wilcoxon rank sum test (with a nominal level of .05) was called for a difficulty check for the selected AT1 items. The number of rejections over 100 replications was noted.

Simulation Case 2

In this case the mean of a_2 values was increased from .49 to 1.23. Thus the distributions of a_1 and a_2 values had the same mean of 1.23 and different SDs with SD_{a1}=0.34 and SD_{a2}=0.11. The purpose for doing this was to explore the sensitivity of DIMTEST to the increased relative potency of the second dimension in terms of the magnitude of item discrimination parameters. Data were simulated with two levels of test length (20, 40), three levels of sample size (500, 1000, 2000), three levels of interability correlation (0, .3, .7), two levels of guessing (0, .2), and two choices of model (compensatory, noncompensatory). The addition of c=0 and ρ=0 cases was intended to examine the behavior of DIMTEST under more extreme circumstances. Each of the 72 combinations of factors was replicated 100 times, resulting in 7200 datasets, with new examinee responses simulated each time. The procedures for generating item responses and for DIMTEST runs were the same as in Case 1.

Simulation Case 3

In this case the means of a_1 and a_2 values remained the same as in case 1, while the SD of a_2 was increased from .11 to .34; thus, the distributions of the a_1 and a_2 values had the same SD of 0.34 and different means with $mean_{a1}$=1.23 and $mean_{a2}$=0.49. This was intended to examine the performance of DIMTEST in the instances with increased relative potency of the second dimension in terms of variability of item discrimination parameters. Comparisons could also be made for rejection rates across the three simulation cases which reflected different distributions of item discrimination parameters. As in Case 2, data were simulated with two levels of test length (20, 40), three levels of sample size (500, 1000, 2000), three levels of interability correlation (0, .3, .7), two levels of guessing (0, .2), and two choices of model (compensatory, noncompensatory). The procedure for generating item responses and for DIMTEST runs remained the same as in previous cases.

Simulation Case 4

In this case both the a_1 and a_2 values had a mean of 1.23 and a SD of 0.34. Interability correlations and guessing were all set to zero. An additional level of correlation between the a_1 and a_2 values (r_{a1a2}=0) was included to identify the effect of correlation of a vectors on the power of DIMTEST. The purpose of this simulation case was to explore the power of DIMTEST in a condition where multidimensionality might be most extreme, with no guessing and zero correlation between dimensions, and with the two dimensions equally potent in terms of both magnitude and variability of item discrimination parameters. Data were simulated with two levels of test length (20, 40), three levels of sample size (500, 1000, 2000), two levels of correlation between a

vectors (0, -0.29), and two choices of model (compensatory, noncompensatory). There were 24 combinations of factors, with each combination replicated 100 times. Again the procedure for item response generation and DIMTEST runs remained the same as outlined in Case 1.

Results

The results are presented separately for each simulation case. Within each case, rejection rates are tabulated separately for the two models. The rows and columns are arranged such that the pattern of numbers is easily captured, thus facilitating the interpretations.

Simulation Case 1

The rejection rates over 100 trials for simulated datasets with varying degrees of test length (N), sample size (S), and interability correlation (ρ) are presented in Table 1 and Table 2 for the noncompensatory model and the compensatory model, respectively. The results are shown for the three significance levels (.01, .05, and .010) with T or T' used. In these tables, each cell (consisting of 3 rows and 6 columns) refers to three datasets with the same level of test length and interability correlation. Within each cell comparisons are made possible for rejection rates across T and T', across α levels, and across sample sizes. The corresponding rows of different cells allow comparison of the effects of test length and interability correlation.

Noncompensatory Model

It can be seen in Table 1 that, as expected, the rejection rates for T' were always higher or at least equal to those for T, and a more liberal α level resulted in higher rejection rates. For datasets with high interability correlation (.7), the rejection rates were all very low. The highest values at the α levels of .01, .05 and .10 were 2%, 4%, and 9%, respectively, using T, and 4% , 7% , and 14%, respectively, using T'. Therefore, the datasets with high interability correlations would all seem to be classified as essentially unidimensional in this case.

Insert Table 1 about here

For datasets with low interability correlations (.3), the results were quite different. For datasets with small (500) and moderate (1000) sample sizes, the rejection rates were all close to nominal levels when T was used and a little higher than nominal levels when T' was used, but even the large values generally did not differ much from the nominal levels. While for datasets with large (2000) sample sizes, the rejection rates were considerably higher than the nominal levels. This was especially true with N=50 and S=2000, where the rejection rates at the significance levels of .01, .05, and .10 were 7%, 17%, and 28%, respectively, using T, and were 12%, 23%, and 33%, respectively, using T'. This would be indicative of some degree of departure from essential unidimensionality. It was also observed that, in some cases, the rejection rates for sample sizes of 500 were higher than those for sample sizes of 1000, which might be due to sampling error. DIMTEST is based on large sample theory; therefore, the results from small samples may be unstable and inaccurate.

It can be concluded that DIMTEST detected some degree of multidimensionality for datasets simulated from the noncompensatory model in Case 1 when the sample size was large and interability correlation was low.

Compensatory Model

From Table 2, it is very clear that the rejections rates were all very low. Across all levels of test length, sample size and interability correlation, the highest rejection rates at the significance levels of .01, .05, .10 were 1%, 5%, and 7%, respectively, using T, and 3%, 8%, and 13%, respectively, using T'. Obviously for all of the datasets simulated from the compensatory model in this Case, the DIMTEST results would imply acceptance of the null hypothesis of essential unidimensionality.

Insert Table 2 about here

Simulation Case 2

For datasets simulated in this case, the rates of correctly rejecting the assumption of essential unidimensionality are presented in Table 3 and Table 4 for the noncompensatory model and compensatory model, respectively.

Noncompensatory Model

Table 3 shows that DIMTEST rejected the hypothesis of essential unidimensionality for datasets simulated from the noncompensatory model under various conditions, and the power was dependent on test length, sample size, interability correlation, and the presence /absence of guessing.

Insert Table 3 about here

The effect of guessing was clearly seen by contrasting the top portion with the bottom portion of Table 3. In general the power decreased when guessing was present. However, under the conditions where both the test length and the sample size were large (N=40, S=2000) and the interability correlation was zero (ρ=0), the power was not greatly affected by guessing. The power increased when test length and sample size increased, and decreased when interability correlation increased, which was in agreement with the results from previous studies based on data simulated from simple structure. For datasets with long tests (N=40) and moderate or large sample sizes (S=1000 or 2000), DIMTEST maintained good power when interability correlation increased from 0 to .3. For all the datasets with test length of 40, sample size of 1000 or 2000, zero guessing, and interability correlation of 0 or .3, the power was extremely high, ranging from 94% to 100% even when a stringent α level of .01 was used. Thus DIMTEST was very powerful for detecting multidimensionality for datasets simulated from the noncompensatory model when test length and sample size were large, interability correlation was low, and no guessing was present.

It was also observed that for various combinations of the following factors, nonzero guessing, short test length, small sample size, and high interability correlation, the rejection rates of DIMTEST dropped to nominal levels. DIMTEST may lack power under these conditions.

Compensatory Model

As shown in Table 4, the rejection rates for the compensatory model were all very low. Across all factors, the highest rejection rates at the α levels of .01, .05, and .10 were 1%, 4%, and 10%, respectively, using T, and were 4%, 10%, and 15%, respectively, using T'. It seems that DIMTEST retained the hypothesis of essential unidimensionality for all the datasets generated from the compensatory model in this case.

Insert Table 4 about here

Simulation Case 3

For datasets simulated in this case, the results for the noncompensatory model and the compensatory model are shown in Table 5 and Table 6, respectively.

Noncompensatory Model

It can be seen from Table 5 that the patterns of effects of test length, sample size, guessing, and interability correlation were similar to those observed in Table 3, but the power was generally lower.

Insert Table 5 about here

With c=.2, the datasets were identified as unidimensional by DIMTEST except when N=40 and ρ=0, where some degree of multidimensionality was detected. With c=0, DIMTEST identified some degree of multidimensionality for all the datasets with N=40 and ρ=0, and the rejection rates were generally higher than the corresponding cases with c=.2. Multidimensionality was also identified for some additional cases with N=20, ρ=0, and S=1000 or 2000, or with N=40, ρ=.3, and S=1000 or 2000.

In general DIMTEST did not show great power for detecting noncompensatory multidimensionality for data generated in this condition. Only in the case where both the test length and sample size were large (N=40, S=2000), the interability correlation was zero (ρ=0), and no guessing was present (c=0), was the power acceptable, with rejection rates at α levels of .01, .05, and .10 being 58%, 74%, and 81%, respectively, using T, and 66%, 78%, and 82%, respectively, using T'.

Compensatory Model

Rejection rates for datasets generated from the compensatory model are shown in Table 6. The power was considerably higher than that for noncompensatory cases under the same conditions. For example, for datasets with N=40, S=1000, and ρ=0, the rejection rates using T' at the α levels of .01, .05, and .10 were 28%, 54%, and 63%, respectively, for the noncompensatory model, and were 75%, 93%, and 95%, respectively, for the compensatory model. Again the same pattern of effects of test length, sample size, interability correlation, and guessing was observed as described before. When guessing was not present, DIMTEST maintained good power for long tests (N=40) and large sample sizes (S=2000) when the interability correlation increased from 0 to .3, which was also observed for the noncompensatory data in Case 2.

Insert Table 6 about here

It can be seen that datasets with ρ=0.7 were uniformly classified as unidimensional, and the same was true for datasets with nonzero guessing and interability correlation of ρ=.3. For datasets with long test lengths (N=40), large sample sizes (S=2000), zero guessing (c=0), and zero to low interability correlations (ρ=0 or ρ=.3), DIMTEST demonstrated good power for detecting compensatory multidimensionality.

Simulation Case 4

For data simulated in this case, the rejection rates are presented in Table 7 and Table 8 for the noncompensatory model and the compensatory model, respectively.

Noncompensatory Model

From Table 7, it can be seen that the rejection rates for the noncompensatory model were comparable to those observed in Case 2 (see Table 3) under the same conditions of test length, sample size, interability correlation, and guessing, and were higher than those observed in Case 3 (see Table 5) under the same conditions. Also, the rejection rates were similar across the two levels of r_{a1a2}. Thus the magnitudes of the discrimination parameters appear to dictate the degree of multidimensionality for datasets simulated form the noncompensatory model, and the performance of DIMTEST was not influenced by the correlation of the a_1 and a_2 values.

Insert Table 7 and Table 8 about here

Compensatory Model

As shown in Table 8, the rejection rates for compensatory model varied significantly across the two levels of r_{a1a2}, with the rejection rates being higher when $r_{a1a2=0}$ than when r_{a1a2} =-.29. This suggests the impact of a_1 and a_2 correlation on the power of DIMTEST for detecting compensatory multidimensionality. Contrasting the portion of Table 8 with r_{a1a2} =-.29 with the left bottom portion of Table 6 reveals that, under these condition, the power was higher when $\text{mean}_{a1} \neq \text{mean}_{a1}$ than when $\text{mean}_{a1} = \text{mean}_{a1}$. Recall that for all the datasets with $\text{mean}_{a1} \neq \text{mean}_{a1}$ and $SD_{a1} \neq SD_{a2}$, and for all the datasets with $\text{mean}_{a1} = \text{mean}_{a1}$ but $SD_{a1} \neq SD_{a2}$, DIMTEST appeared to have limited power for detecting compensatory multidimensionality. This result imples that there is an interaction effect for power between the magnitude and the variability of item discrimination parameters.

Discussion

This study provided preliminary results for the performance of DIMTEST for detecting multidimensionality with data simulated from both compensatory and noncompensatory models under a latent structure that all items in a test were influenced by the same two abilities. The datasets simulated in case 1 were intended to reflect real test data in terms of descriptive statistics and classical item characteristics. The Way et. al study (1988) found that for data simulated as in Case 1, the unidimensional IRT procedures yielded biased estimates of item and ability parameters. Therefore it was of interest to know what DIMTEST would conclude about essential dimensionality for data simulated this way. As the results showed, DIMTEST did identify some degree of departure from essential unidimensionality for datasets simulated from the noncompensatory model when sample size was large and interability correlation was low, although the rejection rates were not very high. On the other hand, for all of the data simulated from the compensatory model, DIMTEST results suggested acceptance of the hypothesis of essential unidimensionality. In general, datasets simulated in this case can be characterized as having one dominant dimension and one minor dimension; therefore, it is not suprising that rejection rates were low in most cases.

In Case 2 through Case 4, data were simulated under various conditions where the relative influence of the second dimension was greater than in Case 1. It should be noted that data simulated in these cases may not be very realistic, and the score distributions resulting from the two models may not be comparable. From a theoretical perspective, however, these cases did allow an assessment and comparison of the performance of DIMTEST for datasets with different distributional characteristics of item discrimination parameters. The results suggested that, for the noncompensatory model, both the magnitude and the variability of a_2 values were related to multidimensionality, with the effect of magnitude dominating over variability for determining multidimensionality. While for the compensatory model, only the variability of the a_2 values seemed to reflect the degree of multidimensionality, and the likelihood of rejecting the hypothesis of essential unidimensionality depended on the interrelation between the a_1 and a_2 values. This finding has some implication for the use of β, proposed by Nandakumar (1991) as a rough index of departure from essential unidimensionality. Since this index was developed for the case of uncorrelated a_1 and a_2 values, it may not be applicable to datasets with nonzero correlation between the a_1 and a_2 values.

In previous studies, DIMTEST had been found to work sufficiently well with data simulated from simple structure type of specification and modeled by compensatory abilities. Given that DIMTEST uses a factor analytic procedure to select AT1 items, intuitively it should work well with this type of latent structure. However, the findings from this study suggest that, for datasets that contained items which were all influenced simultaneously by multiple abilities, the noncompensatory model seemed to be the better approach for modeling the truly non-unidimensional item responses. On the other hand, it might be argued that, the DIMTEST procedure lacks power rather than the compensatory model yields more unidimensional-like data. To address this issue, a comparative study would be necessary so that the number of dimensions for this type of data could be tested using other approaches for assessing unidimensionality.

With respect to the effects of test length, sample size, interability correlation, and guessing, findings across the two models were consistent to those for datasets simulated based on simple structure. Generally speaking, in the situations where DIMTEST identified multidimensionality, the power increased when test length and sample size increased, and when interability correlation decreased. In addition, the power decreased when guessing was present, except when both test length and sample size were large (N=40, S=2000). For datasets with various combinations of short test length (N=20), small sample size (S=500), high correlation between dimensions (ρ=.7), and nonzero guessing (c=.2), DIMTEST appeared to have less power.

This study should foster further research on DIMTEST and other methods for assessing unidimensionality using data simulated from an alternative model and a nonsimple structure type of specification. Although intended to be comprehensive in terms of factors involved, a question remained unanswered, which is, are there monotonic relationships between the power of DIMTEST and the degree of relative magnitude and/or variability for the two discrimination vectors? Future simulation studies with more systematic variations on these factors are needed to address this question.

References

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of multidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement, 9, 37-48.

De Champlain, A., & Gressaroli, M. E. (1991, April). Assessing test dimensionality using an index based on nonlinear factor analysis. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Doody-Bogan, E. N., & Yen, W. M. (1983). Detecting multidimensionality and examining its effects on vertical equating with the three parameter logistic model. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. Applied Psychological Measurement, 29, 1-14.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. Journal of Educational Measurement, 28, 99-117.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. Applied Psychological Measurement. 17, 29-38.

Nandakumar, R., & Stout, W. F. (1993). Refinement of Stout's procedure for assessing latent trait dimensionality. Journal of Educational Statistics, 18, 41-68.

Stout, W. F. (1987): A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. Psychometrika, 55, 293-326.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), Proceeding of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and non-compensatory two-dimensional data on unidimensional IRT estimates. Applied Psychological Measurement, 12, 239-252.

Table 1. Rejection Rates (%) for Noncompensatory Model with $Mean_{a1}$=1.23, $Mean_{a2}$=0.49, SD_{a1}=0.34, SD_{a2}=0.11, r_{a1a2}=-0.29

		ρ=0.3						ρ=0.7					
		α=0.01		α=0.05		α=0.10		α=0.01		α=0.05		α=0.10	
N	S	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'
20	500	0	1	1	7	8	12	0	0	1	3	5	9
	1000	0	1	4	8	10	11	0	0	2	5	6	9
	2000	2	4	9	11	12	19	0	0	0	5	9	11
40	500	3	3	4	12	10	19	2	3	4	7	7	9
	1000	1	3	3	6	8	10	0	1	2	4	5	10
	2000	1	4	10	13	14	20	0	0	1	2	3	6
50	500	0	3	5	12	12	18	1	3	4	6	5	14
	1000	0	2	3	9	11	19	2	4	4	7	8	12
	2000	7	12	17	23	28	33	0	2	3	5	6	10

Notes. N: Test Length

S: Sample Size

ρ: Interability Correlation

Table 2. Rejection Rates (%) for Compensatory Model with $Mean_{a1}=1.23$, $Mean_{a2}=0.49$, $SD_{a1}=0.34$, $SD_{a2}=0.11$, $r_{a1a2}=-0.29$

		ρ=0.3						ρ=0.7					
		α=0.01		α=0.05		α=0.10		α=0.01		α=0.05		α=0.10	
N	S	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'
20	500	0	1	1	5	4	7	0	0	1	3	3	10
	1000	0	0	0	2	3	8	0	0	1	1	2	3
	2000	1	1	1	2	2	2	0	0	0	0	0	1
40	500	1	2	3	5	5	13	0	1	1	4	4	6
	1000	0	0	0	1	2	4	0	1	4	5	5	5
	2000	0	0	0	0	0	0	0	0	0	1	1	1
50	500	1	3	3	8	7	12	1	3	5	7	7	12
	1000	0	1	2	5	3	8	0	0	0	6	6	8
	2000	0	0	0	0	0	1	0	0	1	1	1	1

Notes. N: Test Length

S: Sample Size

ρ: Interability Correlation

Table 3. Rejection Rates (%) for Noncompensatory Model with $Mean_{a1}=Mean_{a2}=1.23$, $SD_{a1}=0.34$, $SD_{a2}=0.11$, $r_{a1a2}=-0.29$

			ρ=0.0						ρ=0.3						ρ=0.7					
			α=0.01		α=0.05		α=0.10		α=0.01		α=0.05		α=0.10		α=0.01		α=0.05		α=0.10	
c	N	S	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'
c=0.2																				
	N=20	S = 500	0	0	2	5	7	12	0	0	1	4	7	13	1	2	4	4	6	9
		S=1000	0	2	4	6	7	11	1	1	1	4	4	6	0	0	1	2	2	6
		S=2000	1	3	11	15	20	22	0	2	3	6	7	10	0	0	0	1	6	12
	N=40	S = 500	20	35	48	58	63	70	11	17	25	31	37	41	0	0	0	4	4	9
		S=1000	52	66	74	82	83	84	17	36	46	51	55	60	1	2	2	3	4	7
		S=2000	92	95	97	99	99	99	48	54	66	71	73	76	1	2	3	5	6	9
c=0																				
	N=20	S = 500	0	1	7	21	25	36	0	0	2	7	10	17	0	0	0	3	3	6
		S=1000	5	15	26	33	38	48	2	6	11	26	29	35	0	0	1	2	4	7
		S=2000	36	59	75	81	82	87	14	32	46	54	59	69	0	0	7	15	17	22
	N=40	S = 500	87	89	95	99	100	100	53	70	79	89	89	89	2	8	12	22	21	30
		S=1000	97	99	100	100	100	100	94	96	97	97	98	98	11	16	21	32	33	38
		S=2000	100	100	100	100	100	100	100	100	100	100	100	100	53	61	61	67	67	72

Notes. N: Test Length

S: Sample Size

ρ: Interability Correlation

Table 4. Rejection Rates (%) for Compensatory Model with $Mean_{a1}=Mean_{a2}=1.23$, $SD_{a1}=0.34$, $SD_{a2}=0.11$, $r_{a1a2}=-0.29$

			ρ=0.0						ρ=0.3						ρ=0.7					
			α=0.01		α=0.05		α=0.10		α=0.01		α=0.05		α=0.10		α=0.01		α=0.05		α=0.10	
c	N	S	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'
c=0.2																				
	N=20	S = 500	0	0	0	5	3	6	0	2	2	4	4	6	0	1	2	3	3	3
		S=1000	0	0	0	1	0	1	0	0	0	1	2	2	0	0	0	2	1	3
		S=2000	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	N=40	S = 500	1	4	4	5	6	7	0	0	1	2	1	4	0	0	0	6	6	10
		S=1000	0	0	0	1	2	3	0	0	0	0	0	1	0	0	0	0	0	2
		S=2000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
c=0																				
	N=20	S = 500	0	0	0	1	1	4	0	0	0	0	1	3	0	0	0	0	0	3
		S=1000	0	1	2	4	4	7	0	0	0	0	1	3	0	0	0	1	1	4
		S=2000	0	1	3	7	10	15	0	0	0	2	3	6	0	0	0	0	0	3
	N=40	S = 500	0	2	4	10	10	15	0	0	2	4	4	8	0	0	0	3	3	7
		S=1000	0	0	2	7	7	9	0	0	0	0	1	2	0	0	0	1	1	1
		S=2000	0	0	1	3	4	10	0	0	0	0	1	2	0	0	0	0	1	7

Notes. N: Test Length
S: Sample Size
ρ: Interability Correlation

Table 5. Rejection Rates (%) for Noncompensatory Model with $Mean_{a1}$=1.23, $Mean_{a2}$=0.49, SD_{a1}=SD_{a2}=0.34, r_{a1a2}=-0.29

			ρ=0.0						ρ=0.3						ρ=0.7					
			α=0.01		α=0.05		α=0.10		α=0.01		α=0.05		α=0.10		α=0.01		α=0.05		α=0.10	
c	N	S	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'
c=0.2																				
	N=20	S = 500	0	0	0	3	3	6	0	2	5	8	10	14	0	2	2	4	6	8
		S=1000	1	1	3	3	5	10	1	1	1	2	3	13	0	0	0	0	1	3
		S=2000	1	2	4	10	10	12	0	1	1	2	4	8	0	0	3	4	4	4
	N=40	S = 500	2	2	4	11	11	14	0	0	2	3	3	9	1	3	3	5	4	9
		S=1000	2	3	6	13	16	21	0	1	4	4	5	8	0	1	1	1	1	3
		S=2000	5	10	15	18	21	24	0	1	3	4	4	5	0	0	0	1	1	3
c=0																				
	N=20	S = 500	1	4	7	9	10	15	0	0	1	3	5	6	0	0	2	6	5	10
		S=1000	1	3	4	10	11	19	0	3	4	7	9	12	0	1	2	4	4	5
		S=2000	5	9	20	27	30	37	3	4	6	11	11	19	1	2	6	11	11	13
	N=40	S = 500	3	14	23	29	32	39	2	5	6	8	9	13	2	2	3	4	4	7
		S=1000	16	28	44	54	57	63	3	4	6	11	11	20	0	0	0	2	2	2
		S=2000	58	66	74	78	81	82	7	11	13	19	21	24	0	0	0	0	0	0

Notes. N: Test Length
S: Sample Size
ρ: Interability Correlation

Table 6. Rejection Rates (%) for Compensatory Model with $Mean_{a1}$=1.23, $Mean_{a2}$=0.49, SD_{a1}=SD_{a2}=0.34, r_{a1a2}=-0.29

			ρ=0.0						ρ=0.3						ρ=0.7					
			α=0.01		α=0.05		α=0.10		α=0.01		α=0.05		α=0.10		α=0.01		α=0.05		α=0.10	
c	N	S	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'
c=0.2																				
	N=20	S = 500	1	1	5	10	11	16	0	0	2	3	4	10	0	0	0	4	5	8
		S=1000	1	2	10	13	16	19	0	0	1	5	7	10	0	0	0	3	5	7
		S=2000	5	10	16	24	28	35	0	0	0	2	3	5	0	0	0	1	3	3
	N=40	S = 500	6	10	13	19	22	27	0	2	3	7	8	12	1	2	2	4	4	4
		S=1000	10	19	30	37	41	47	1	2	4	6	7	10	0	0	0	1	1	2
		S=2000	32	40	43	50	52	55	0	0	0	0	0	1	0	0	0	0	1	7
c=0																				
	N=20	S = 500	3	8	14	20	22	28	1	4	6	13	16	21	0	0	0	3	3	4
		S=1000	10	23	32	46	48	62	5	9	15	18	25	36	0	0	1	1	1	3
		S=2000	46	62	80	83	83	85	14	33	49	60	69	73	0	0	0	1	2	5
	N=40	S = 500	16	35	48	62	64	75	12	25	34	50	53	63	0	0	0	2	2	9
		S=1000	67	75	86	93	93	95	37	56	64	75	82	84	1	1	1	2	2	5
		S=2000	89	91	93	97	97	98	80	90	92	94	95	97	0	0	0	0	1	7

Notes. N: Test Length

S: Sample Size

ρ: Interability Correlation

Table 7. Rejection Rates (%) for Noncompensatory Model with $\rho=0$, $c=0$, $Mean_{a1}=Mean_{a2}=1.23$, $SD_{a1}=SD_{a2}=0.34$

		$r_{a1a2}=-0.29$						$r_{a1a2}=0$					
		$\alpha=0.01$		$\alpha=0.05$		$\alpha=0.10$		$\alpha=0.01$		$\alpha=0.05$		$\alpha=0.10$	
N	S	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'
20	500	0	3	6	14	16	23	1	3	7	13	16	24
	1000	5	14	23	35	40	53	8	17	25	39	40	52
	2000	42	51	62	71	72	79	26	41	58	72	78	82
40	500	84	89	95	97	98	98	88	92	94	95	95	96
	1000	99	99	100	100	100	100	100	100	100	100	100	100
	2000	100	100	100	100	100	100	100	100	100	100	100	100

Notes. N: Test Length

S: Sample Size

r_{a1a2}: Correlation between *a* Vectors

Table 8. Rejection Rates (%) for Concompensatory Model with ρ=0, c=0, $Mean_{a1}$=$Mean_{a2}$=1.23, SD_{a1}=SD_{a2}=0.34

		r_{a1a2}=-0.29						r_{a1a2}=0					
		α=0.01		α=0.05		α=0.10		α=0.01		α=0.05		α=0.10	
N	S	T	T'	T	T'	T	T'	T	T'	T	T'	T	T'
20	500	0	1	5	12	16	26	0	0	0	4	5	7
	1000	4	5	7	9	13	24	0	0	0	3	3	5
	2000	8	11	15	21	24	32	1	2	9	16	17	21
40	500	10	25	32	45	48	65	4	6	6	12	13	27
	1000	37	59	66	82	83	85	12	23	34	43	43	52
	2000	69	78	78	83	84	88	28	42	45	56	56	61

Notes. N: Test Length

S: Sample Size

r_{a1a2}: Correlation between a Vectors

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM031634

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Detecting compensatory and Noncompensatory Multidimensionality Using DIMTEST	
Author(s): Hui Deng and Timothy N. Ansley	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY ______ Sample ______ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY ______ Sample ______ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY ______ Sample ______ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ [ ]	Level 2A ↑ [X]	Level 2B ↑ [ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: Hui Deng	Printed Name/Position/Title: Hui Deng	
Organization/Address: 346 LC the University of Iowa Iowa City, IA 52242	Telephone: (319) 335-5259	FAX: (319) 335-6038
	E-Mail Address: hui-deng@uiowa.edu	Date:

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20772
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 2/2000)